

Paket Biplot Biasa dan Kekar dengan Pemrograman Fungsional Mathematica Berbasis GUI

N.K. KUTHA ARDANA DAN SISWADI

Departemen Matematika FMIPA IPB
Jl. Meranti, Kampus IPB Darmaga, Bogor 16680, Indonesia

Abstrak

Paket biplot sebagai alat eksplorasi dan visualisasi data peubah ganda dikembangkan dengan teknik pemrograman fungsional *Mathematica* berbasis GUI (*Graphical User Interface*). Paket ini dapat digunakan untuk melakukan eksplorasi data peubah ganda, baik tanpa pencilan maupun dengan pencilan. Analisis biplot biasa untuk kasus data tanpa pencilan menggunakan penguraian nilai singular matriks data dengan norma L_2 , sedangkan biplot kekar untuk kasus matriks data yang mengandung pencilan dilakukan dengan memberikan bobot pada setiap baris matriks data dengan pendugaan- M kekar. Konfigurasi peubah-objek ditampilkan di dalam ruang dua dimensi pada berbagai koefisien $\alpha \in [0, 1]$ yang dapat diubah secara interaktif, disertai ukuran kesesuaian matriks pendekatan, dan berbagai hasil komputasi numerik lainnya. Ilustrasi numerik diberikan untuk melihat tampilan kedua jenis biplot yang dihasilkan.

Kata kunci: biplot, biplot kekar, penguraian nilai singular, pendugaan- M , pemrograman fungsional.

Pengantar

Metode biplot memegang peranan penting sebagai suatu teknik eksplorasi data peubah ganda yang dapat memvisualisasikan secara serempak n baris objek dan p kolom peubah suatu matriks data ke dalam suatu ruang berdimensi rendah (Hardle dan Simar, 2003). Dalam grafik biplot, vektor-vektor baris yang mewakili objek ditumpangtindihkan dengan vektor-vektor kolom yang mewakili peubah sehingga dimungkinkan untuk memperoleh informasi tentang: i) kedekatan antar objek, ii) keragaman dan korelasi antar peubah, maupun iii) keterkaitan antara objek dan peubah.

Sebagai alat peraga data peubah ganda yang praktis, implementasi metode ini terus dikembangkan. Lipkovich dan Smith (2002) menerapkan metode ini dalam kaitan dengan analisis statistika peubah ganda seperti Analisis Komponen Utama, Analisis Korespondensi, Penskalaan Multidimensional. Biplot biasa dengan sistem perintah juga telah diintegrasikan ke dalam beberapa program paket Statistika seperti SAS, R, Stata. Sejalan dengan makin berkembangnya teknik komputasi dengan sistem aljabar komputer (SAK), biplot (biasa) telah diimplementasikan ke dalam SAK Mathematica dengan output berupa fungsi perintah disertai beberapa opsi (Kutha dan Siswadi, 2005). Sistem perintah seperti ini dapat menyulitkan pengguna secara umum karena banyaknya perintah dan opsi-opsi yang mesti diingat. Pada tulisan ini dibahas implementasi metode biplot biasa dan kekar dengan pemrograman fungsional Mathematica berbasis GUI sehingga paket yang dihasilkan akan sangat memudahkan pengguna.

Biplot Biasa

Misalkan sebuah matriks data X^* terdiri atas n objek (sampel pengamatan) untuk masing-masing p peubah dan lakukan koreksi terhadap rataannya atau bakukan dalam hal peubah memiliki ukuran atau skala berbeda, sehingga diperoleh matriks data terkoreksi X . Pada kasus umum atau tanpa pencilan, visualisasi ke- n objek dan p peubah matriks data X^* secara serempak pada ruang berdimensi rendah (umumnya 2) dapat dilakukan dengan metode biplot biasa. Metode ini didasarkan pada meminimumkan jumlah kuadrat terkecil

$$\sum_{i=1}^n \sum_{j=1}^p (x_{ij}^* - x_{ij})^2 \quad (1)$$

Solusi dari masalah meminimuman ini dapat dinyatakan sebagai penguraian nilai singular (PNS) dari matriks data X dengan pangkat $r \leq p \leq n$ (Gabriel, 1971).

$$X = U \Lambda V' \quad (2)$$

dengan $U(n \times r)$ dan $V(p \times r)$ merupakan matriks ortonormal kolom ($U'U = V'V = I_r$) dan $\Lambda(r \times r) = \text{diag}(\sqrt{\lambda_1}, \sqrt{\lambda_2}, \dots, \sqrt{\lambda_r})$ yang bersifat $\sqrt{\lambda_1} \geq \sqrt{\lambda_2} \geq \dots \geq \sqrt{\lambda_r}$. Unsur λ_i , $i = 1, 2, \dots, r$ adalah nilai eigen $X'X$ atau XX' , dan $\sqrt{\lambda_i}$ disebut nilai singular (Mardia et al., 1979). Matriks V adalah matriks yang kolom-kolomnya terdiri atas vektor-vektor eigen v_i yang berpadanan dengan nilai eigen λ_i dari matriks $X'X$. Kolom-kolom matriks U terdiri atas vektor-vektor eigen yang berpadanan dengan nilai-nilai eigen matriks XX' , $U = (Xv_1/\sqrt{\lambda_1}, Xv_2/\sqrt{\lambda_2}, \dots, Xv_r/\sqrt{\lambda_r})$

Definisikan $\Lambda^\alpha = \text{diag}(\sqrt{\lambda_1}^\alpha, \sqrt{\lambda_2}^\alpha, \dots, \sqrt{\lambda_r}^\alpha)$, $\alpha \in [0, 1]$, dan misalkan $G = U \Lambda^\alpha$, $H = V \Lambda^{1-\alpha}$ maka (2) dapat dituliskan menjadi

$$\begin{aligned} X &= U \Lambda V' \\ &= (U \Lambda^\alpha) (\Lambda^{1-\alpha} V') \\ &= G H' \end{aligned} \quad (3)$$

Dengan demikian unsur ke- (i, j) dari matriks data X ($n \times p$) dapat dinyatakan sebagai

$$x_{ij} = g_i' h_j \quad (4)$$

dengan g_i' , $i = 1, 2, \dots, n$, dan h_j' , $j = 1, 2, \dots, p$ berturut-turut adalah vektor-vektor baris matriks G dan H masing-masing dengan r unsur (Jolliffe, 2002). Di sini, n baris dari G berkaitan dengan baris-baris dari X , dan p baris dari H berhubungan dengan kolom-kolom dari matriks X .

Persamaan (4) juga bermakna bahwa nilai x_{ij} diwakili oleh proyeksi g_i pada h_j . Nilai x_{ij} bertanda positif (nilai pengamatan semula, x_{ij}^* > nilai rata-rata \bar{x}_j) bila sudut antara vektor g_i dan h_j , $\theta \in [0, \pi/2)$, dan bertanda negatif (nilai $x_{ij}^* < \bar{x}_j$) bila $\theta \in (\pi/2, \pi]$. Nilai x_{ij} mendekati nol (nilai pengamatan semula, x_{ij}^* mendekati rata-rata \bar{x}_j) bila g_i dan h_j saling ortogonal. Jadi, posisi relatif titik-titik g_i dan h_j pada biplot akan memberikan informasi tentang besaran objek ke- i pada peubah ke- j (Kutha dan Siswadi, 2005).

Untuk $r = 2$, matriks data X dapat digambarkan secara pasti pada ruang berdimensi dua. Bila $r > 2$ dan matriks data X ingin digambarkan pada ruang berdimensi $k < r$, dapat dilakukan suatu pendekatan terbaik dengan suatu matriks berpangkat k (umumnya $k = 2$). Dalam hal ini, X pada persamaan (3) didekati dengan

$$\hat{X} = GH' = (U_k \Lambda_k^\alpha) (\Lambda_k^{1-\alpha} V_k') \quad (5)$$

■ $\alpha = 0$ (biplot GH)

Pemilihan $\alpha = 0$ pada (3) berakibat $X'X = (n-1)S$, dengan S menyatakan matriks ragam-koragam S . Ini berarti bahwa panjang vektor h_j pada biplot menggambarkan keragaman x_j , nilai kosinus sudut antara h_i dan h_j merepresentasikan korelasi antara peubah x_i dan x_j . Hal lainnya, dari vektor-vektor g , dapat ditunjukkan bahwa jarak Euclid antara g_i dan g_j pada biplot sebanding dengan jarak Mahalanobis antara vektor x_i dan x_j (Jolliffe, 2002).

■ $\alpha = 1$ (biplot JK)

Jika $\alpha = 1$ pada (3) maka $G = \Lambda$, $H = V$. Akibatnya $XX' = GG'$. Pada kasus ini, jarak Euclid antara g_i dan g_j pada biplot juga menggambarkan jarak Euclid antara x_i dan x_j .

■ $\alpha = 0.5$ (biplot simetrik)

Dengan mengambil $\alpha = 0.5$ pada (3), interpretasi seperti pada biplot GH dan JK tidak berlaku, namun posisi relatif g_i dan g_j pada biplot masih mencerminkan informasi tentang besaran objek ke- i pada peubah ke- j ($x_{ij} = g_i' h_j$). Bentuk plot yang dihasilkan berada di antara biplot GH dan JH karena pemberian bobot/skala yang seimbang antara koordinat-koordinat g_i dan h_j (Lipkovich dan Smith, 2002).

Biplot Kekar

Biplot biasa sensitif terhadap keberadaan data pencilan sehingga disarankan menggunakan metode lain yang lebih tahan terhadap data pencilan. Biplot kekar (*robust*) di sini diperoleh dengan memberikan bobot pada baris matriks data. Makin ekstrem tingkat pencilan sebuah titik data, makin kecil bobot yang diberikan (Daigle and Rivest, 2002). Dengan demikian, masalah meminimuman (1) menjadi

$$\sum_{i=1}^n \sum_{j=1}^p \hat{w}_i (x_{ij} - x_{ij})^2 \quad (6)$$

Nilai pembobot baris ke- i , w_i , diperoleh melalui proses iteratif sistem persamaan taklinear berikut (Maronna, 1976).

$$\frac{\sum_{i=1}^n \sqrt{w_i} x_i}{\sqrt{w_i}} = \mu \quad (7)$$

$$\sum_{i=1}^n \sqrt{w_i} (x_i - \mu)(x_i - \mu)' = \Sigma \quad (8)$$

$$w_i = \frac{h((x_i - \mu)' \Sigma^{-1} (x_i - \mu))}{(n-1)C} \quad (9)$$

$\hat{\mu}$ dan $\hat{\Sigma}$ didefinisikan sebagai rata-rata kekar dan matriks koragam kekar. Fungsi h merupakan fungsi pembobot yang bersifat menurun pada $(0, \infty)$ dan C suatu konstanta penormalan. Proses iteratif tersebut merupakan pendugaan- M peubah ganda vektor lokasi μ dan matriks sebaran Σ . Pada proses ini rata-rata dan koragam diboboti dengan bobot pada masing-masing titik data bergantung pada jarak terhadap lokasi pendugaan.

Dalam hal ini digunakan fungsi pembobot Huber

$$h(t) = \min \left(t, p - \frac{\sqrt{p}}{2} \right) \quad (10)$$

yang akan menurunkan bobot sebuah titik data yang jarak Mahalanobisnya ke penduga pusat lebih besar dari $p - \sqrt{p}/2$. Konstanta penormalan C dipilih untuk membuat pendugaan- M dari matriks sebaran $\hat{\Sigma}$ sebagai penduga konsisten dari matriks koragam bila contoh diambil dari populasi p -peubah normal. Untuk fungsi h yang dipilih pada (10), $\hat{\Sigma}$ memenuhi sifat ini bila

$$C = P \left[\chi^2_{p+2} \leq p - \frac{\sqrt{p}}{2} \right] + \frac{p - \sqrt{p}/2}{p} P \left[\chi^2_{p+2} > p - \frac{\sqrt{p}}{2} \right] \quad (11)$$

(Maronna, 1976).

Setelah diperoleh matriks diagonal W berukuran $n \times n$ dengan unsur diagonal utama \hat{w}_i , solusi masalah meminimuman (6) dapat dinyatakan sebagai PNS dari

$$W^{1/2} X = U^w \Lambda^w (V^w)' \quad (12)$$

(Jolliffe, 2002), dengan $U^w (n \times r)$ dan $V^w (p \times r)$ merupakan matriks ortonormal kolom dan $\Lambda^w (r \times r) = \text{diag}(\sqrt{\lambda_1^w}, \sqrt{\lambda_2^w}, \dots, \sqrt{\lambda_r^w})$ yang bersifat $\sqrt{\lambda_1^w} \geq \sqrt{\lambda_2^w} \geq \dots \geq \sqrt{\lambda_r^w}$. Unsur λ_i^w , $i = 1, 2, \dots, r$ adalah nilai eigen pendugaan- M dari matriks koragam kekar $\hat{\Sigma}$, dan V^w adalah matriks vektor-vektor eigen dari $\hat{\Sigma}$.

Di sini X adalah matriks data yang telah terkoreksi terhadap mediannya. Dalam hal peubah memiliki ukuran atau skala berbeda, dilakukan pembakuan dengan matriks diagonal V yang merupakan invers dari matriks ragam kekar, $1/\hat{\Sigma}_{jj}$.

Peminimuman X menghasilkan

$$\hat{X} = W^{-1/2} U_k^w \Lambda_k^w V_k'^w \quad (13)$$

dengan U_k^w dan V_k^w diambil dari k kolom pertama matriks U^w dan V^w , Λ_k^w adalah matriks diagonal berunsur k nilai singular pertama. Seperti pada (3), misalkan $G = (W^{-1/2} U_k^w \Lambda_k^{w\alpha})$, dan $H = V_k^w \Lambda_k^{w1-\alpha}$, maka selanjutnya dapat dikonstruksi biplot kekar dengan berbagai nilai $\alpha \in [0, 1]$ (Daigle and Rivest, 2002).

Ukuran Kesesuaian Biplot

Gabriel (2002) memberikan berbagai ukuran kesesuaian (GF, *goodness of fit*) antara suatu matriks dengan matriks pendekatannya dalam analisis biplot. Matriks data X didekati dengan matriks GH' , koragam dan korelasi antar peubah $X'X$ didekati dengan HH' , kemiripan objek XX' didekati dengan GG' . Makin besar (mendekati 100%) koefisien GF, makin sesuai matriks pendekatannya merepresentasikan matriks awalnya, dan karenanya makin layak hasil analisis biplot digunakan untuk penarikan kesimpulan.

Bentuk ukuran pendekatan yang digunakan adalah sebagai berikut:

$$1. \text{Kesesuaian data : GF}(X, GH') = \frac{\text{tr}^2(X' GH')}{\text{tr}(X' X) \text{tr}(HG' GH')} \quad (14)$$

$$2. \text{Kesesuaian peubah : GF}(X' X, HH') = \frac{\text{tr}^2(X' XHH')}{\text{tr}(X' XX') \text{tr}(HH' HH')} \quad (15)$$

$$3. \text{Kesesuaian peubah : GF}(XX', GG') = \frac{\text{tr}^2(XX' GG')}{\text{tr}(XX' XX') \text{tr}(GG' GG')} \quad (16)$$

Implementasi dengan Mathematica

Sistem aljabar komputer *Mathematica* sangat andal dalam pemrograman fungsional yang menghasilkan program yang ringkas dengan waktu eksekusi yang lebih cepat dibandingkan pemrograman prosedural. Pemrograman fungsional pada *Mathematica* dicirikan dengan proses pemetaan fungsi terhadap suatu list (kumpulan ekspresi, data) dengan menggunakan fungsi-fungsi dasar seperti Map, Apply, Nest, FixedPoint. Dimulai dari versi 6, *Mathematica* menyediakan perintah Manipulate yang efisien dalam memvisualisasikan suatu perintah fungsi secara interaktif (Ruuskeepaa, 2009). Fasilitas ini digunakan untuk mengimplementasikan biplot biasa dan kekar.

PNS suatu matriks data m dilakukan dengan menggunakan fungsi SingularValueDecomposition[m]. Fungsi SVD2[m] dibuat untuk memodifikasi fungsi SingularValueDecomposition[m] dengan penyesuaian tanda bagi vektor-vektor eigennya sehingga menghasilkan plot yang lebih mudah penginterpretasiannya.

Fungsi muSigmaWeight[m, iter] menghasilkan pendugaan kekar-M pada biplot kekar bagi rata-rata kekar $\hat{\mu}$, koragam kekar $\hat{\Sigma}$, dan pembobot \hat{w} melalui proses iterasi dengan perintah FixedPoint dengan nilai awal $\mu_0 = \text{Mean}[X]$, $\Sigma_0 = \text{Covariance}[X]$, dan w_0 menggunakan fungsi pembobot Huber dengan $\mu = \text{Median}[X]$ dan $\Sigma_0 = \text{Covariance}[X]$ pada (9). Proses iterasi ini berhenti setelah paling banyak terjadi iter kali.

Nilai koefisien $\alpha \in [0, 1]$ dapat diatur dengan menggeser slidernya. Dalam kasus tampilan vektor-vektor peubah kurang seimbang (terlalu panjang/pendek) terhadap posisi objek, slider stretch dapat digunakan untuk mengaturnya. Kotak centang Standardize (Robust Standardize) dapat ditandai pada kasus peubah-peubah perlu dibakukan karena tidak memiliki ukuran atau skala yang sama.

Plot yang dihasilkan berupa tebaran koordinat objek yang ditumpangtindihkan dengan plot vektor peubah berupa garis lurus berarah ditampilkan pada tab Biplot. Tab Numerics menampilkan keluaran komputasi numerik. Hasil komputasi numerik juga dapat ditampilkan dalam sel untuk keperluan analisis lebih lanjut, atau dalam format tabel.

Ilustrasi Numerik

- Berikut adalah fungsi untuk membangkitkan matriks data $n \times p$ penambahan/pengurangan bilangan acak dengan sebaran normal

```
generateData[n_, p_] :=
Module[{μ = 1, α = Range[-Floor[(n) / 2], Floor[(n - 1) / 2]],
  β = Range[-Floor[(p) / 2], Floor[(p - 1) / 2]]},
Table[μ + α[[i]] + β[[j]] +
  RandomReal[NormalDistribution[0, 0.125^0.5]],
{i, Length[α]}, {j, Length[β]}]
]
```

- Dibangkitkan matriks data 10×5

```
mat = generateData[10, 5] // N; mat // MatrixForm
```

```
{-6.34958 -5.22402 -4.59554 -3.07252 -1.80932
-5.28709 -4.04698 -3.27758 -2.50785 -1.46534
-3.77602 -2.82051 -1.74188 -1.42317 -0.0253592
-2.59797 -2.53895 -1.27308 0.366862 1.04569
-2.20013 -1.14629 -0.313258 1.27683 2.25785
-1.1377 0.163722 0.480993 1.32562 2.84133
0.585748 1.07231 1.88263 2.8765 3.44784
1.31714 1.57635 2.28382 3.81136 4.45254
1.75281 2.93927 3.82736 4.66366 5.15138
3.36782 3.46671 5.93841 5.81178 6.96132}
```

- Biplot biasa

```
Biplot[mat]
```

$$X \approx (U L^{\alpha_1})(L^{\alpha_2} V') = G H'$$

Standardize ☐

Object's Label Automatic

Variable's Label Automatic

AspectRatio 1

StretchH ☐ 1

$\alpha_2 = 1 - \alpha_1$ ☒

$\alpha_2 = \alpha_1$ ☐

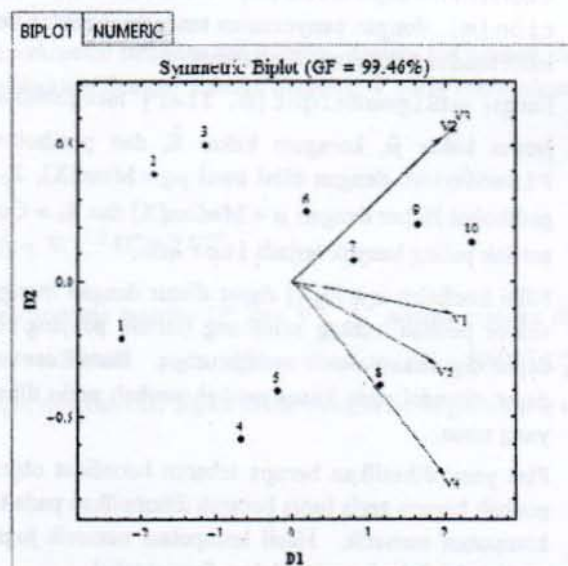
α_1 0.5

α_2 0.5

View Numerical Results in Cell

View Tables in Cell

About



Biplot kekar

WeightedBiplot[mat]

$$X \approx (U L^{\alpha 1})(L^{\alpha 2} V') = G H'$$

Robust Standardize ☐

Object's Label: Automatic

Variable's Label: Automatic

AspectRatio: 1

StretchH: ☐ = 1

$\alpha 2 = 1 - \alpha 1$ ☒

$\alpha 2 = \alpha 1$ ☐

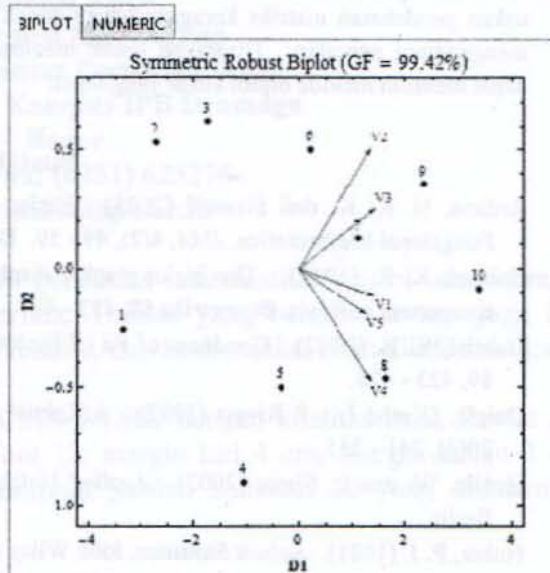
$\alpha 1$: ☐ = 0.5

$\alpha 2$: ☐ = 0.5

View Numerical Results in Cell

View Tables in Cell

About



Tampak bahwa kedua biplot menghasilkan konfigurasi peubah-objek yang berbeda. Pada biplot biasa peubah v_2 hampir berimpit dengan peubah v_3 . Ini berbeda pada tampilan biplot kekar, karena adanya pembobotan yang diberikan pada masing-masing objek.

Berikut adalah hasil iterasi bagi ratahan kekar $\hat{\mu}$, koragam kekar $\hat{\Sigma}$, dan pembobot \hat{w} . Perhatikan objek pencilan baris ke-10 mendapat bobot terkecil.

$$\begin{pmatrix} -1.36276 \\ -0.552947 \\ 0.364201 \\ 1.39434 \\ 2.35649 \end{pmatrix}$$

$$\begin{pmatrix} 20.7482 & 19.3468 & 20.9141 & 19.6077 & 18.495 \\ 19.3468 & 18.5064 & 19.6692 & 18.3208 & 17.3829 \\ 20.9141 & 19.6692 & 21.5633 & 19.8223 & 18.84 \\ 19.6077 & 18.3208 & 19.8223 & 18.8214 & 17.6655 \\ 18.495 & 17.3829 & 18.84 & 17.6655 & 16.8272 \end{pmatrix}$$

$$\begin{pmatrix} 0.207435 \\ 0.22933 \\ 0.202833 \\ 0.266032 \\ 0.256087 \\ 0.327426 \\ 0.214879 \\ 0.254994 \\ 0.309623 \\ 0.174725 \end{pmatrix}$$

Simpulan dan Saran

Metode biplot biasa dan biplot kekar telah diimplementasikan ke dalam paket sistem aljabar komputer Mathematica dengan teknik pemrograman fungsional berbasis GUI. Biplot biasa cocok untuk eksplorasi data peubah ganda tanpa pencilan, sedangkan biplot kekar yang menggunakan pendekatan matriks koragam kekar cocok untuk kasus eksplorasi peubah ganda yang mengandung pencilan. Disarankan untuk membandingkan berbagai kriteria kekekaran untuk dapat memilih metode biplot kekar yang tepat.

Pustaka

- Ardana, N. K. K. dan Siswadi (2005). Biplot dan Implementasinya dengan Pemrograman Fungsional Mathematica. *JMA*, 4(2), 49 - 59. Departemen Matematika FMIPA-IPB. Bogor.
- Gabriel, K. R. (1971). The biplot-graphic display of matrices with application to principal component analysis. *Biometrika* 58, 453 - 467.
- Gabriel, K. R. (2002). Goodness of Fit of Biplots and Correspondence Analysis. *Biometrika* 89, 423 - 426.
- Daigle, G. and L. P. Rivest (1992). A Robust Biplot. *The Canadian Journal of Statistics*, 20(3), 241 - 255.
- Hardle, W. dan L. Simar (2003). *Applied Multivariate Statistical Analysis*. Springer-Verlag, Berlin.
- Huber, P. J. (1981). *Robust Statistics*. John Wiley and Sons, USA.
- Johnson, R. A. dan D. W. Wichern (2002). *Applied Multivariate Statistical Analysis*. 5th ed. Prentice-Hall, Inc., USA.
- Jolliffe, I.T. (2002). *Principal Component Analysis*. 2nd ed. Springer-Verlag, NewYork.
- Lipkovich, I. and E. P. Smith (2002). Biplot and Singular Value Decomposition Macros for Excel®. *Journal of Statistical Software* 7(5), 1-15.
- Mardia, K. V., J. T. Kent, dan J. M. Bibby (1979). *Multivariate Analysis*. Academic Press, London.
- Maronna, R.A. (1976). Robust M-estimators of multivariate location and scatter. *Ann. Statist.*, 4, 51-67.
- Maronna, R.A., R.D. Martin, dan V.J. Yohai (2006). *Robust Statistics: Theory and Methods*. John Wiley and Sons, USA.
- Ruskeepaa, H. (2009). *Mathematica Navigator: Mathematics, Statistics, and Graphics*. Elsevier, UK.